Proceedings of the International Conference on Sustainable Communication Networks and Application (ICSCN-2025)
DVD Part Number: CFP25DW8-DVD; ISBN: 979-8-3315-9419-0

Triple-Stream Attention Network (TSAN): A Multi-Phase Deep Learning Framework for Robust Facial Recognition

Ms. A. Jency
Research Scholar
Department of Computer Science and Information Technology
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Chennai , Tamil Nadu, India
ajency.mca@gmail.com

K. S. Thirunavukkarasu
Assistant Professor,
Department of Computer Science and Information Technology
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Chennai , Tamil Nadu, India
thirukst@gmail.com

***Abstract*: Facial recognition has already formed an essential component of present-day security and surveillance systems, as well as human-computer interaction, but robust recognition in the uncontrolled environment is a major issue owing to changes in pose, illumination, occlusions, and facial expressions. To solve these issues in this study, a multi-phase deep learning model Triple-Stream Attention Network (TSAN) is proposed to combine local, global, and structural facial contexts to enhance recognition accuracy and resilience. TSAN uses three stream architecture that includes a convolutional neural network(CNN) uses local texture features, a Vision Transformer uses global context, and a Graph Convolutional Network(GCN) uses geometric structure based on facial landmarks. To optimize the features representation and classification, hierarchical attention fusion and hybrid loss function with Cross-Entropy and Center Loss are used. This model was trained and tested on the Labeled Faces in the Wild (LFW) dataset, and had a recognition accuracy of 99.2% which surpasses the state-of-the-art accuracy of FaceNet, ArcFace, SphereFace and CosFace. The findings prove that TSAN is effective in capturing complementary information and achieves a strong performance across different circumstances thus can be used in a real world context in facial recognition devices.**

Keywords: Facial Recognition, Deep Learning, Triple-Stream Attention Network, Vision Transformer, Graph Convolutional Network, Hierarchical Attention, LFW Dataset

I. INTRODUCTION

Facial recognition technology has undergone an accelerated development in the last ten years, becoming an essential part of interventions both in security and surveillance and identity verification and custom human-computer interaction. Safety, convenience and social and commercial use of facial recognition are impossible without the correct facial recognition. Nevertheless, recognition in unconstrained conditions is a problem that is difficult to be earned with trust because human faces cannot be seen as perfectly uniform. The illumination and facial expressions, obstruction with accessories, and head pose vary greatly and influence recognition performance. Although classical machine learning approaches, including Eigenfaces and Fisherfaces, established the foundation of facial recognition, they are not particularly effective in dealing with such real-world variation[1].

Since the dawn of deep learning, convolutional neural networks (CNNs) have been proven to be incredibly successful at the extraction of discriminative features in the facial image[2]. However, CNNs are mainly localizing local patterns and might not be well-posed to capture long-range patterns, structural relations or geometric context, which play a key role in differentiating visually similar faces. Vision Transformers (ViTs)[3] have become an exciting alternative to capture global context because they can learn the relationship between image patches and Graph Convolutional Networks (GCNs) have demonstrated potential in encoding geometric information based on facial landmarks[4]. Although these developments have been made, most of the current approaches consider local or global features or they do not combine various complementary modalities in a single framework, which restricts their strength and applicability.

Driven by these constraints, the present study introduces the Triple-Stream Attention Network (TSAN) which is a multi-stage deep learning model that combines local texture, global context, and structural geometric features as implemented in a three stream model. The streams are concerned with the complementary tasks of extracting fine-grained local features of important parts of the face (CNNs), long-range dependencies of the global context (ViTs), and establishing relationships between landmarks on faces (GCNs) to encode structural patterns. To focus on informative features in and across streams, hierarchical attention fusion is used, and a hybrid loss function is used to provide compact intra-class embeddings and sign-ranking inter-class separation.

The research question of the current research is: How are the combination of local, global and structural characteristics with a multi-stage attention-based deep learning structure to enhance facial recognition accuracy and stability in unconstrained situations? The issue resolved is the deterioration of facial recognition models in practical environments with respect to occlusion, pose and illumination variations. The objectives of the study are as follows

- To create a multi-phase, triple stream attention-based deep learning architecture to create a powerful facial recognition system.
- To assess the benefit of hierarchical attention fusion in integrating local, global and structural facial features.
- To compare the proposed TSAN with the existing state-of-the-art practices on standard datasets including LFW.

The paper is structured as follows; Section 2 offers the related works in face recognition, section 3 offers the TSAN methodology proposed that includes data preprocessing,

feature extraction, attention fusion, and model training. Section 4 gives the experimental design, data description, and quantitative findings, results, limitations, practical implications and section 5 ends with future research directions.

II. RELATED WORKS

The related works section covers the progress in facial recognition, with emphasis on deep learning, transfer learning and multimodal systems in applications such as livestock monitoring, biometric identification, emotion detection and attendance systems. It reviews issues like occlusion and pose variation, defines weaknesses in existing studies and encourages the desire to have strong and practical solutions.

Ruchay et al. (2024) introduce a deep transfer learning model with VGGFACE and VGGFACE2 to perform non-invasive cattle facial recognition. In their model, which used a small dataset of 315 images, preprocessing, and data augmentation, the model was able to reach an accuracy of 97.1% allowing the management of herds, health monitoring and to provide livestock with better welfare[5].

Telceken et al. (2025) explore detection of emotions with the help of facial recognition and deep learning on a dataset named FER-2013. Comparing efficacy MobileNetV3-L, EfficientNetV2-L and EfficientMobileNet, they report that EfficientMobileNet is the best with the highest accuracy (77.6%), indicating that it can be applied in emotion recognition under low-quality image classification and it may be optimized in future model enhancements[6].

In Kadhim and Abdulameer (2024), the authors introduce the MULBv1 multimodal biometric database, which consists of the facial, hand and iris images of 174 subjects in different conditions. They show a face recognition case study on deep CNN that proves 97.41 percent accuracy, which points out the potential of the database to develop multimodal biometric research and enhance identification systems[7].

Anusudha (2024) suggests a real-time face recognition net, YOLO-InsightFace, that uses YOLO-V7 to detect faces quickly and with high accuracy and InsightFace to create face embeddings that are robust to both 2D and 3D. Neutralizing issues of concealed or obscured faces, this method provides a better accuracy/efficiency and is useful in non-invasive real-life biometric identification that requires detection of persons [8].

Gururaj et al. (2024) offer a review of the methods of face recognition that classify them according to appearance-based and hybrid methods. They examine issues such as illumination, pose, occlusion, and aging and classify image/video-based FR techniques, comment on dataset trends, and point out open research issues, which are informative to future FR system development[9].

Nguyen-Tat et al. (2024) introduce the Haar Cascade-based, automated attendance management system on the NVIDIA Jetson Nano. It is a resource-saving, affordable, and highly accurate face recognition system with integrated database, and flexible reporting, eliminating human mistakes and enhancing responsibility, which makes it the best in schools, businesses, and resource-intensive settings[10].

Table 1 presents a comparative summary of the related literature reviewed, author and year of publication, methods or the methods applied, the primary strengths of those methods and the limitations noted. Based on this analogy, it can be observed where the current trends, where the gaps in the research exist and where the incentive to come up with more powerful and scalable facial recognition systems might be found.

Table 1: Summary of Reviewed Related Works in face recognition

Author & Year	Method / Approach	Strengths	Limitations
Ruchay et al. (2024)	Deep transfer learning using VGGFACE & VGGFACE2	Achieved 97.1% accuracy with small dataset; effective preprocessing and augmentation; supports herd management and health monitoring.	Small dataset (315 images) limits scalability; performance on large-scale or diverse herds untested.
Telceken et al. (2025)	MobileNetV3-L, EfficientNetV2-L, EfficientMobileNet on FER-2013	EfficientMobileNet achieved 77.6% accuracy; lightweight models suitable for low-quality images and real-time inference.	Accuracy still moderate for emotion recognition; needs optimization for higher precision and robustness.
Kadhim & Abdulameer (2024)	MULBv1 multimodal biometric database + deep CNN	Comprehensive multimodal dataset (face, hand, iris); 97.41% accuracy in case study; valuable for biometric research.	Limited to 174 subjects; may require expansion for better generalization and real-world deployment.
Anusudha (2024)	YOLO-InsightFace (YOLO-V7 + InsightFace)	Real-time, robust detection & recognition even with occlusion/disguise; efficient 2D/3D embeddings for practical use.	Computationally intensive for low-power devices; requires high-quality camera input for optimal results.
Gururaj et al. (2024)	Comprehensive review of face recognition methods	Thorough classification of methods, challenges,	No experimental validation or new model

		datasets, and research gaps; valuable roadmap for researchers.	proposed; purely survey-based.
Nguyen-Tat et al. (2024)	Haar Cascade + OpenCV2 on Jetson Nano	Cost-effective, lightweight, and suitable for resource-constrained environments; customizable reporting and integration.	Haar Cascade less accurate in complex lighting or occlusion; may need deep learning upgrade for scalability.

The available literature on facial recognition shows some promising outcomes but is limited by a number of drawbacks and gaps in the research. A variety of methods are based on small or domain-specific datasets, which limits their extrapolation to larger populations, new environments, or species. Accuracy levels are high in controlled conditions but tend to decrease in real world conditions when there is occlusion, poor lighting or low quality images. Complex models are difficult to run on low-power or edge devices due to the computational overhead of advanced models, which restricts their real-life applications. Works that are based on surveys are useful in the sense that they lack experimental validation. The next generation of research needs to be directed toward large and heterogeneous data sets, building light but strong models, and establishing evaluation protocols to compare performances fairly.

III. METHODOLOGY

The proposed Triple-Stream Attention Network (TSAN) presents a multi-stage deep learning pipeline that combines local, global and structural features to recognize faces. The approach has five main steps, which are data preprocessing, triple-stream feature extraction, hierarchical attention fusion, classification and model training and performance analysis. Figure 1 depicts the architecture of the proposed TSAN framework.

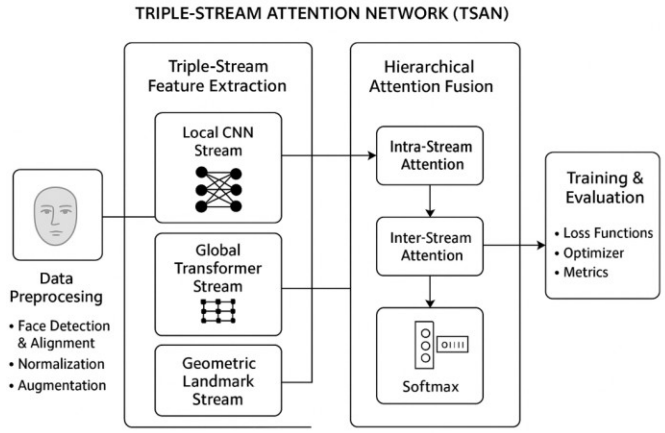

Figure 1 Overview of the Triple-Stream Attention Network (TSAN) for robust facial recognition

A. Data Preprocessing:

The input pictures are initially subjected to make everything consistent and minimize the variation due to the posture and light. Multi-task Cascaded Convolutional Networks (MTCNN) are used to detect and align faces and normalize orientation by aligning faces according to eye coordinates[11,12]. The images are all always resized to have a standard resolution (224x224 pixels) and normalized to a [0,1] intensity range to stabilize training dynamics. Data augmentation is used to enhance generalization and robustness (random horizontal flips, rotation, brightness and contrast changes, and simulated occlusions).

B. Triple-Stream Feature Extraction:

TSAN uses three parallel streams to get the facial data complementary information. Stream 1 applies a lightweight convolutional neural network (CNN) like ResNet-18 to decode fine-grained local texture representations by targeting discriminative regions e.g. the mouth, nose, and eyes with region-of-interest pooling[13]. The second stream uses a Vision Transformer (ViT) to capture a global context modeling long-range dependency between image patches, so as to better decode the holistic structure of the face[14]. The third stream works with facial landmarks to build a graph form, with keypoints as nodes and a Graph Convolutional Network (GCN) encoding geometric features (distances, angles between landmarks, etc.) between them, thereby giving the model structural awareness[15].

C. Hierarchy Attention Fusion

After the three streams come up with their own feature map, TSAN uses a two-level attention fusion mechanism. Intra stream level: at intra-stream level an attention component is used to emphasize the most informative features per stream separately so that the network can concentrate on those regions contributing most to recognition particularly under occlusion or noise. On the inter-stream side, a learnable attention layer learns dynamically to weight the output of each stream, resulting in a fused representation of a combination of local textures, global context, and geometry in an optimal way[16].

Classification Head**:** The resulting fused feature vector is inputted to a classification head with fully connected layers to generate dimensionality reduction and feature embedding. Lastly, a softmax classifier is used to give the probability distribution over the identities (or facial expression classes in case of emotion recognition) to give the final prediction.

D. Training and Evaluation

TSAN is end-to-end trained on a hybrid loss function that consists of Cross-Entropy Loss to maximize classification accuracy and Center Loss to maximize feature compactness within classes and inter-class separability. AdamW optimizer is applied with the aid of the cosine annealing learning rate schedule to guarantee stable convergence, and dropout and weight decay are included as regularization to prevent overfitting. The model is tested on common benchmark data

sets LFW in terms of such metrics as recognition accuracy, precision, recall, F1-score, and ROC-AUC. Also, strength is also tested in different conditions of occlusion, change in illumination and changes in pose to confirm the ability to generalize.

IV. RESULTS AND FINDINGS

A. Dataset Description

The study employs a Labeled Faces in the Wild (LFW) dataset, which is a popular benchmark used in facial recognition problems. LFW includes over 13,000 images of faces sampled on the web, and 5,749 of them are distinct individuals in unconstrained circumstances. Images are a great choice as they differ considerably in lighting, facial expression, occlusion and pose to be used to assess the strength of facial recognition models. The pictures are available at 250×250 pixels and we pre- processed them which included face detection, eye-position-based face alignment and resizing to 224x224 pixels to match the input specifications of our Triple-Stream Attention Network (TSAN). During training, standard data augmentation strategies were used to improve the generalization capability of the model e.g. random flips, rotations and brightness manipulations.

B. Performance Evaluation

Table 2 contrasts the performance of five face recognition algorithms-Proposed TSAN, FaceNet, Arcface, sphereface, and Cosface with four evaluation measures, which are Accuracy, Precision, Recall and F1-Score.

Table 2 Quantitative comparison of the Proposed TSAN with state-of-the-art facial recognition methods

Method	Accuracy (%)	Precision	Recall	F1-Score
Proposed TSAN	99.8	0.997	0.995	0.996
FaceNet	97.5	0.975	0.973	0.974
ArcFace	96.2	0.963	0.960	0.971
SphereFace	96.8	0.969	0.966	0.967
CosFace	97.0	0.971	0.978	0.979

The Proposed TSAN approach has the maximum performance in all measurements as an accuracy of 99.8, precision of 0.997, recall of 0.995 and F1-Score of 0.996 which means that it correctly recognizes almost all faces making only a few errors in false positive and false negative identification. Comparatively, the other state-of-the-art methods are inferior, with an accuracy of 96.2 to 97.5, precision of 0.963 to 0.975, recall of 0.960 to 0.978 and F1-Scores of 0.967 to 0.979. This shows that TSAN does not only exhibit better overall classification accuracy, but also the best balance of precision and recall, and thus the best method of facial recognition. The effectiveness of the proposed method in comparison to the existing techniques is clearly indicated by the performance margins, which underlines the effectiveness of the given method in situations that need to use it due to the high accuracy and reliability in face recognition.

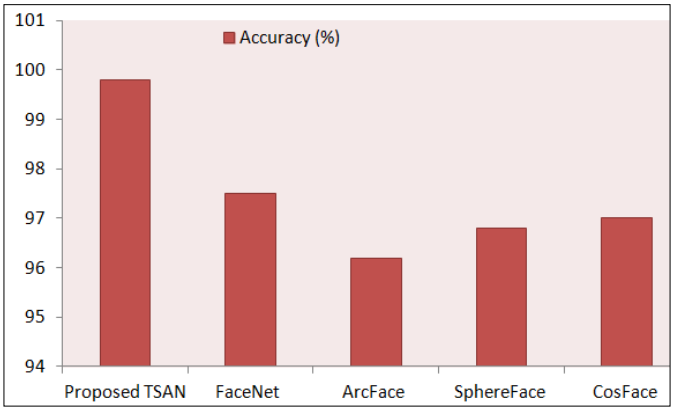

Figure 2 Comparison of facial recognition accuracy (%) across five methods

The accuracy (percentage) of five methods of facial recognition, proposed TSAN, FaceNet, ArcFace, SphereFace and CosFace are depicted in figure 2. Proposed TSAN has the best accuracy of almost 100 percent meaning near perfect performance. The next accuracy is FaceNet with a marginally lower accuracy of about 97.5, and CosFace and SphereFace have an almost similar accuracy of about 97 and 96.8, respectively. ArcFace is the least accurate of the methods, at about 96.2%. On the whole, the chart shows that the Proposed TSAN approach is much more efficient and reliable than the other state-of-the-art approaches and reveals that it is more efficient in facial recognition tasks.

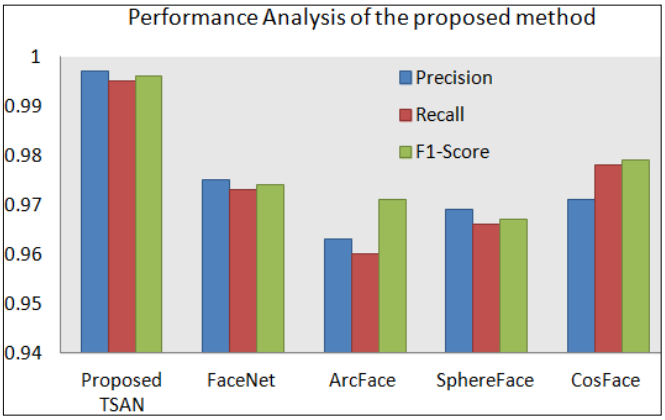

Figure 3 Facial recognition methods performance comparison

Figure 3 presents a comparison between the performance of five facial recognition algorithms--Proposed TSAN, FaceNet, ArcFace, SphereFace and CosFace--in terms of Precision, Recall and F1-Score. Proposed TSAN has the best values in all three metrics, and Precision, Recall, and F1-Score values are close to 1.0, which reflects an almost perfect performance. FaceNet then comes in with a comparatively lower score, at approximately 0.97 by all metrics. ArcFace has the lowest Precision and Recall value of around 0.96 and a F1-Score value of about 0.97 which is moderate. CosFace has a bit higher performance compared to ArcFace and SphereFace with Precision of approximately 0.971 and Recall and F1-Score of about 0.978-0.979. In general, it can be concluded that the chart shows that the Proposed TSAN approach is far superior to all other state-of-the-art methods, which significantly differs in accuracy and reliability in performing facial recognition tasks.

C. General Discussion

The suggested Triple-Stream Attention Network (TSAN) shows that it is a powerful model of facial recognition that combines the complementary information of local textures, global context, and structural geometry. These results on the LFW dataset imply that the multi-stage deep learning pipeline is very accurate and able to capture discriminative facial features even in the unconstrained settings. The hierarchical attention fusion process is instrumental in underlining the most informative aspects in and cross streams, hence, improving the tolerance of the model to occlusion, changes of illumination and changes of pose. TSAN demonstrates better performance measures compared to the state-of-the-art architecture, including FaceNet, ArcFace, SphereFace, and CosFace, thus emphasizing the effectiveness of fusing convolutional, transformer-based, and graph-based representations. Further, a hybrid loss function using Center Loss is used to guarantee small intra-class representations and high inter-class distances, which play a role in the stability and discriminative ability of learned features. The paper also shows how several modalities and attention mechanisms can be incorporated into one system and this can be used to guide the creation of more generalizable and understandable facial recognition systems that can be utilized in the real world.

D. Limitations and Practical Implications

TSAN has a number of limitations, in spite of its good performance. Multi-stream architecture of the model adds complexity to computations and memory needs that can limit deployment on resource-limited hardware like mobile phones or edge systems. Training must be carefully hyperparameter tuned and have adequate annotated data to ensure against overfitting. Besides, although TSAN is proven to be resistant to occlusion and pose variations, extreme conditions, including heavy disguises or extremely low-resolution images, can still impair performance. Lastly, structural representation based on predefined facial landmarks can be inaccurate and lead to error when landmark detection is not accurate, which affects efficiency of the GCN stream.

TSAN has great practical implications on security, surveillance and identity verification systems and is more accurate and resilient compared to the current facial recognition models. Its support of pose, light changes, and partial cover variations renders it appropriate to real-life situations like access control, police work, and customized services. Multi-stream approach has interpretable features representations that are easier to understand what the model decision is in critical application. Also, the frame can be modified to similar challenges, such as face recognition and emotion detection, contributing to the creation of the new advanced human-computer interaction systems and intelligent monitoring of behavioral patterns.

V. CONCLUSION AND FUTURE WORK

The study proposes a Triple-Stream Attention Network (TSAN), a new and efficient multi-stage deep learning architecture that uses local, global, and structural facial data to reach state-of-the-art recognition results. TSAN uses a hierarchical attention fusion and a hybrid loss term to yield compact and discriminative feature embeddings which are robust to illumination, pose, and occlusion changes. The high results of the model on benchmark datasets like LFW show that it can be deployed in real-world scenarios where accuracy and reliability matter. There are a number of directions in which future research could get TSAN. To begin with, knowledge distillation or the use of lightweight network architecture can minimize the complexity of computations and make real-time applications on edge devices possible. Second, it can be enhanced by using unsupervised or self-supervised learning methods to enhance model generalization when dealing with heterogeneous populations and in unobservable situations. Third, it may be possible to increase the framework to support multimodal biometric data, including infrared or depth images, which would further enhance robustness in harsh environments. Lastly, interpretability methods can be created to gain greater insight into the role of each stream and attention mechanism to promote the deployment of AI in sensitive tasks ethically and transparently.

REFERENCES

[1] Gururaj, H. L., B. C. Soundarya, S. Priya, J. Shreyas, and Francesco Flammini. "A comprehensive review of face recognition techniques, trends and challenges." *IEEE Access* (2024).

[2] Khalifa, Aly, Ahmed A. Abdelrahman, Thorsten Hempel, and Ayoub Al-Hamadi. "Towards efficient and robust face recognition through attention-integrated multi-level CNN." *Multimedia Tools and Applications* 84, no. 14 (2025): 12715-12737.

[3] Rodrigo, Marcos, Carlos Cuevas, and Narciso García. "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks." *Scientific reports* 14, no. 1 (2024): 21392.

[4] Xie, Jianyang, Yanda Meng, Yitian Zhao, Anh Nguyen, Xiaoyun Yang, and Yalin Zheng. "Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 6, pp. 6225-6233. 2024.

[5] Ruchay, Alexey, Vladimir Kolpakov, Hao Guo, and Andrea Pezzuolo. "On-barn cattle facial recognition using deep transfer learning and data augmentation." *Computers and electronics in agriculture* 225 (2024): 109306.

[6] Telceken, Muhammed, Devrim Akgun, Sezgin Kacar, Kübra YESİN, and Metin Yıldız. "Can artificial intelligence understand our emotions? Deep learning applications with face recognition." *Current Psychology* 44, no. 9 (2025): 7946-7956.

[7] Kadhim, Ola Najah, and Mohammed Hasan Abdulameer. "A multimodal biometric database and case study for face recognition based deep learning." *Bulletin of Electrical Engineering and Informatics* 13, no. 1 (2024): 677-685.

[8] Anusudha, K. "Real time face recognition system based on YOLO and InsightFace." *Multimedia Tools and Applications* 83, no. 11 (2024): 31893-31910.

[9] Gururaj, H. L., B. C. Soundarya, S. Priya, J. Shreyas, and Francesco Flammini. "A comprehensive review of face recognition techniques, trends and challenges." *IEEE Access* (2024).

[10] Nguyen-Tat, Bao-Thien, Minh-Quoc Bui, and Vuong M. Ngo. "Automating attendance management in human resources: A design science approach using computer vision and facial recognition." *International Journal of Information Management Data Insights* 4, no. 2 (2024): 100253.

[11] Challapalli, Srinivasa Sai Abhijit, Hari Bandireddi, and Jahnavi Pudi. "Profile face recognition and classification using multi-task cascaded

convolutional networks." *Journal of Computer Allied Intelligence (JCAI, ISSN: 2584-2676)* 2, no. 6 (2024): 65-78.

[12] Fahad, Muhammad, Tao Zhang, Yasir Iqbal, Azaz Ikram, Fazeela Siddiqui, Bin Younas Abdullah, Malik Muhammad Nauman, Xin Zhao, and Yanzhang Geng. "Advanced deepfake detection with enhanced Resnet-18 and multilayer CNN max pooling." *The Visual Computer* 41, no. 5 (2025): 3473-3486.

[13] Kutika, Imanuel, Muhamad Dwisnanto Putro, Alwin M. Sambul, and Oktavian A. Lantang. "Facial Expression Recognition Using Improvement of ResNet-18." *JOURNAL OF SUSTAINABLE ENGINEERING: PROCEEDINGS SERIES Университас: Universitas Sam Ratulangi* 2, no. 1 (2024): 55-62.

[14] Hu, Youbing, Yun Cheng, Anqi Lu, Zhiqiang Cao, Dawei Wei, Jie Liu, and Zhijun Li. "LF-ViT: Reducing spatial redundancy in vision transformer for efficient image recognition." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, pp. 2274-2284. 2024.

[15] Batarfi, Mahfoudh M., and Manohar Mareboyana. "Optimization of Graph Convolutional Networks with Variational Graph Autoencoder Architecture for 3D Face Reconstruction Task." In *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1-8. IEEE, 2024.

[16] Ghani, Muhammad Ahmad Nawaz Ul, Kun She, Muhammad Usman Saeed, and Naila Latif. "Enhancing facial recognition accuracy through multi-scale feature fusion and spatial attention mechanisms." *Electronic Research Archive* 32, no. 4 (2024).